NCBI GEO

Archive for gene expression, epigenomic & functional genomic data sets

Gene Expression Omnibus

ncbi.nlm.nih.gov/geo/

Gene Expression Omnibus archives and freely distributes comprehensive gene expression, epigenomic and functional genomic data sets. Sequence- and array-based data are accepted. Tools help users search, analyze, visualize and download studies and gene expression profiles.

over 100,000 studies
comprising
2.6 million assays
from 4,000 species

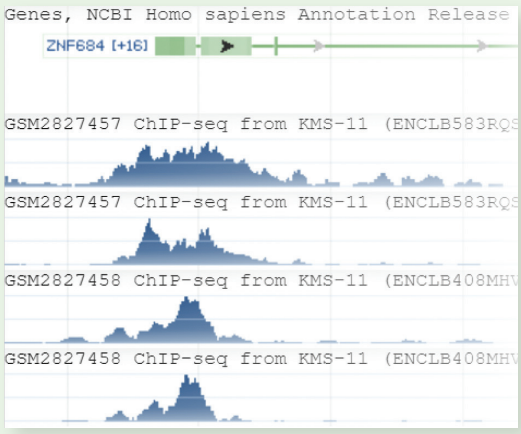
Study types included in GEO

RNA-seq · ChIP-seq · miRNA-seq
single-cell-seq · RNA-array · miRNA-array
ChIP-array · arrayCGH · ATAC-seq · methyl-seq · HiC-seq and more...

On the web

Search data
go.usa.gov/xUS6k

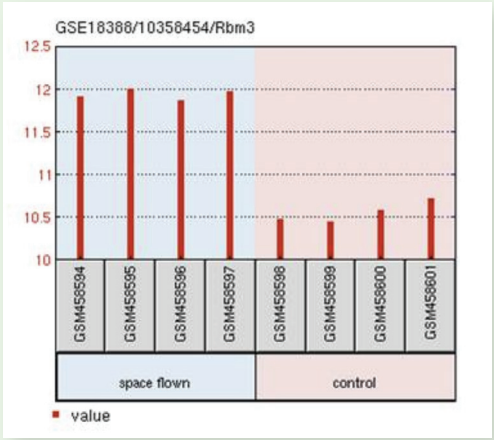
View data with tracks on Genome Data Viewer
ncbi.nlm.nih.gov/gds/?term=track[filter]



Download data
ncbi.nlm.nih.gov/geo/info/download.html

Analyze with GEO2R
go.usa.gov/xUS6X

Use GEO2R to identify genes that are differentially expressed across a study.



Watch our GEO2R tutorial
bit.ly/GEO2R-tutorial

NIH U.S. National Library of Medicine
National Center for Biotechnology Information

How do I submit to NCBI GEO?

GEO accepts many categories of high-throughput functional genomic data, including all array-based applications and some high-throughput sequencing data.

Submissions include:

- raw data files
- metadata spreadsheet
- processed data files

ncbi.nlm.nih.gov/geo/info/submission.html

INSTRUCTIONS

Step 1: Check that GEO accepts your data type
Step 2: Create a GEO account
Step 3: Gather your raw data files
Step 4: Gather your processed data files
Step 5: Fill in the GEO Metadata Template (see below)
Step 6: Transfer the entire data set to GEO by FTP
Step 7: Notify GEO once the FTP transfer is complete
Step 8: A GEO curator reviews your submission and issues your accession numbers, typically within 5 business days

ncbi.nlm.nih.gov/geo/info/submission.html

GEO facts

- Your GEO accession numbers can be cited in manuscripts discussing the data.
- You can update your GEO records at any time.
- Your GEO records can be held private until your manuscript is published.
- You can allow reviewers anonymous access to your private GEO records.
- No need to submit separately to SRA, GEO submits raw data to SRA for you.

ncbi.nlm.nih.gov/geo/info/faq.html

For help, contact us at geo@ncbi.nlm.nih.gov

Metadata Template

Describe the overall study, including title, summary and contributors.

List and describe each of the biological samples under investigation. Examples include organism, tissue, age.

List and describe experimental protocols. Examples include growth, treatment, library construction protocols.

List and describe data processing steps. Examples include base calls, alignment, peak calls, normalization.

List processed file names and types. Examples of processed files include BIGWIG, FPKM, CHP, TXT.

List raw file names for example, FASTQ, BAM, or CEL files.

SERIES

title	Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.
summary	We report the application of single-molecule-based sequencing technology for high-th
overall design	Examination of 2 different histone modifications in 2 cell types.
contributor	John,B,Goode
contributor	Bradley,Smith

SAMPLES

Sample name	title	source name	organism	characteristics
Sample 1	H3K4me2_ChIPSeq	Neural progenitor cells	Mus musculus	ES-derived neur
Sample 2	H3K4me1_ChIPSeq	Neural progenitor cells	Mus musculus	ES-derived neur
Sample 3	input DNA	Neural progenitor cells	Mus musculus	ES-derived neur

PROTOCOLS

growth protocol	ES cell–derived NS cells were routinely generated by re-plating d 7 adherent neural di
treatment protocol	
extract protocol	Lysates were clarified from sonicated nuclei and histone-DNA complexes were isolate
library construction protocol	Libraries were prepared according to Illumina's instructions accompanying the DNA S
library strategy	ChIP-Seq

DATA PROCESSING PIPELINE

data processing step	Basecalls performed using CASAVA version 1.4. ChIP-seq reads were aligned to the r
data processing step	Data were filtered using the following specifications…
data processing step	peaks were called using PeaksFind version 2.2 with the following setting: ChIP thresh
genome build	mm9
processed data files format and content	wig files were generated using …; Scores represent …

PROCESSED DATA FILES

file name	file type	file checksum
H3K4me2.peaks.wig	wig	95cf1d1fa509d871b2ef0bb9fd734c3d
H3K4me1.peaks.wig	wig	8ec6ee3cce10b970e5bfea4e35cdb231
H3K4me2.b.peaks.wig	wig	f8fcd650914ff1a733956d6d06e8b543

RAW FILES

file name
080716_BI-EAS46_0001_209DH_L1.fastq

NIH U.S. National Library of Medicine
National Center for Biotechnology Information